Testing Parametric versus Semiparametric Modelling in Generalized Linear Models

Härdle, W.K.; Mammen, E.; Müller, M.D.

Publication date:
1996

Testing Parametric versus Semiparametric Modelling in Generalized Linear Models*

Wolfgang HÄRDLE

Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät
Humboldt-Universität zu Berlin, Germany

Enno MAMMEN

Institut für Angewandte Mathematik
Ruprecht-Karls-Universität Heidelberg, Germany

Marlene MÜLLER

Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät
Humboldt-Universität zu Berlin, Germany

June 5, 1996

We consider a generalized partially linear model $E(Y|X,T) = G\{X^T\beta + m(T)\}$ where G is a known function, β is an unknown parameter vector, and m is an unknown function. The paper introduces a test statistic which allows to decide between a parametric and a semiparametric model: (i) m is linear, i.e. $m(t) = t^T\gamma$ for a parameter vector γ, (ii) m is a smooth (nonlinear) function. Under linearity (i) it is shown that the test statistic is asymptotically normal. Moreover, for the case of binary responses, it is proved that the bootstrap works asymptotically. Simulations suggest that (in small samples) bootstrap outperforms the calculation of critical values from the normal approximation. The practical performance of the test is shown in applications to data on East–West German migration and credit scoring.

*The research for this paper was supported by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" at Humboldt University, Berlin (Germany). The work of M. Müller was supported in part by CentER, Tilburg University (The Netherlands). We thank Michael C. Burda for helpful discussion and comments on the economic applications.

1 Introduction

In the analysis of discrete response variables one often models the expected value of the response as a nonlinear monotone function of a linear combination of the explanatory variables. Examples are Probit or Logit models where the nonlinear (link) function is the cumulative distribution function of a normal respectively logistic distribution, see McCullagh and Nelder (1989). Then the so-called *generalized linear model* has the form

$$E(Y|Z) = G(Z^T\theta) \tag{1.1}$$

with a known monotone function G and an unknown parameter θ. The model (1.1) combines computational feasibility (especially for discrete covariates) with good interpretability of the "index" $Z^T\theta$ and therefore has found wide application in all fields of applied statistics, see e.g. Fahrmeir and Tutz (1994), Maddala (1983). However, for some applications it may be argued that the assumption of linearity in (1.1) is too restrictive. Indeed it may be not even clear if the relationship between the influential variables and the response is monotone. A more complex relationship (allowing also for nonmonotone dependence) is given by the following semiparametric *generalized partially linear model*

$$E(Y|Z) = G\{X^T\beta + m(T)\} \tag{1.2}$$

where $Z = (X, T)$ is a split of Z into two components X and T, β is an unknown parameter and m is an unknown smooth function. For a discussion of model (1.2) and for further references, see Severini and Staniswalis (1994).

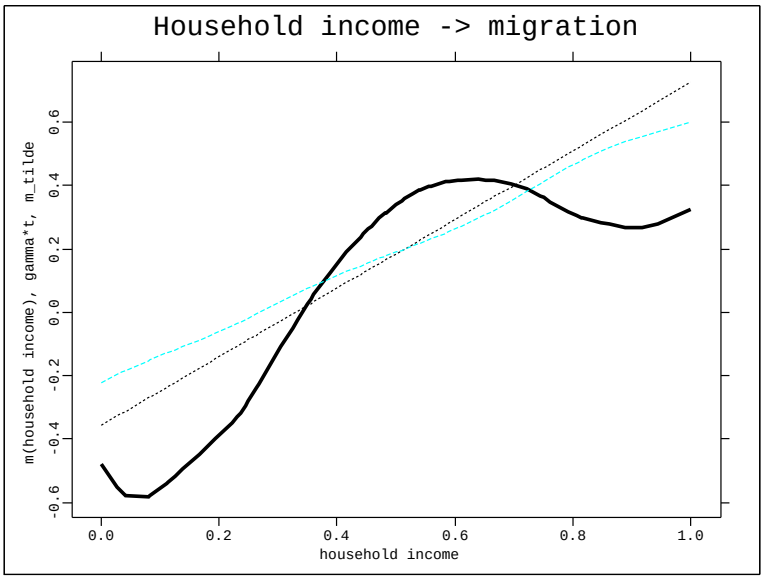

Figure 1: The influence $m(t)$ of household income (transformed to $[0, 1]$) on migration intention. Nonparametric fit (thick black line), linear fit (thin black dashed line), and "biased" parametric estimate $\widetilde{m}$ (see (2.9), thin grey dashed line), $n = 402$.

As an example for a possible nonlinear dependence consider a model on East–West German migration in 1991 (data from the German Socio-Economic Panel for Mecklenburg-Vorpommern, a Land of the Federal State of Germany, GSOEP, 1991). The dependent variable is binary with $Y = 1$ (intention to move) or $Y = 0$ (intention to stay). As an explanatory variable serves besides some socioeconomic factors $X =$ (age, sex, friends in west, city size, unemployment) the variable $T =$ household income. Figure 1 shows a fit of the function m in the semiparametric model (1.2) using a logistic link function $G(u) = 1/\{1 + \exp(-u)\}$. The estimated function is clearly nonlinear and shows a saturation in the intention to migrate for higher income housholds. The question is of course, whether the observed nonlinearity is significant.

In this paper we will discuss tests of the parametric hypothesis (1.1),

$$m(t) = t^T\gamma \quad \text{for a vector } \gamma, \tag{1.3}$$

versus the semiparametric alternative (1.2). This test gives a first indication if new shapes observed in nonparametric fits of m are significant. Furthermore, the proposed test complements the work of Severini and Staniswalis (1994), who consider estimation under model (1.2). With identity link this model has been also analysed by Green (1987), Speckman (1983) and Robinson (1988). For another related model see Carroll, Fan, Gijbels and Wand (1995). Most of the literature in this semiparametric context though was devoted to estimation and not to testing.

The next Section 2 introduces estimators of m, γ and β. These estimators will be used in the construction of the test statistics. The test and its asymptotic properties for the case of a binary response are discussed in Section 3. Section 4 reports on a small simulation study, the application to the migration example and another example on credit scoring. Remarks on the computation of the test statistics and proofs of our results are given in the appendix.

2 Estimation in the Parametric and in the Semiparametric Model

For the estimation of the parametric component β and the nonparametric component m we follow the approach of Severini and Staniswalis (1994). The method is based on quasi-likelihood estimation. The quasi-likelihood function is defined as

$$Q(\mu; y) = \int_{\mu}^{y} \frac{(s-y)}{V(s)}\, ds$$

where μ is the (conditional) expectation of Y, i.e. $\mu = G\{X^T\beta + m(T)\}$. It is assumed here that the conditional variance of Y is $\sigma^2 V(\mu)$ where σ is an unknown scale parameter

and V is a known function. Quasi-likelihood functions are motivated by exponential families. Note that the maximum likelihood estimate $\widehat{\theta}$, based on an i.i.d. sample $Y_1, ..., Y_n$ from an exponential family, is given by

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta} Q(\mu_i; Y_i) = 0.$$

In our model the quasi-likelihood function is given as

$$\mathcal{L}(m, \beta) = \sum_{i=1}^{n} Q(\mu_i; Y_i) \tag{2.1}$$

where $(Y_1, X_1, T_1), \ldots, (Y_n, X_n, T_n)$ is a sample of independent observations and $\mu_i = G\{X_i^T \beta + m(T_i)\}$. The parameter β is supposed to lie in $B \subset I\!R^p$. The covariates X_i, T_i are $I\!R^p$ and $I\!R^q$ valued. We assume that the response variable Y_i is real valued. Multidimensional responses can be treated similarly.

For the estimation of the nonparametric component m we make use of the following smoothed quasi-likelihood

$$\mathcal{L}^S(m(.), \beta) = \int \sum_{i=1}^{n} K_h(t - T_i)\, Q[G\{X_i^T \beta + m(t)\}; Y_i]\, dt, \tag{2.2}$$

where $K_h(u) = (h_1 \cdot \ldots \cdot h_q)^{-1} K(h_1^{-1} u_1, ..., h_q^{-1} u_q)$ is a kernel (defined on $I\!R^q$) with bandwidth (vector) $h = (h_1, ..., h_q)$. Following Severini and Staniswalis (1994), Severini and Wong (1992) we put for $\beta \in B$

$$\widehat{m}_\beta = \arg\min_m \mathcal{L}^S(m, \beta), \tag{2.3}$$

$$\widehat{\beta} = \arg\min_\beta \mathcal{L}(\widehat{m}_\beta, \beta), \tag{2.4}$$

$$\widehat{m} = \widehat{m}_{\widehat{\beta}}. \tag{2.5}$$

In (2.3) minimization runs over functions $m(.)$. Therefore the value $\eta = \widehat{m}_\beta(t)$ is defined as the minimizer of $\sum_{i=1}^n K_h(t - T_i)\, Q[G\{X_i^T \beta + \eta\}; Y_i]$, see (2.2). Without loss of generality we always assume that the constant vector is not contained in the design space. An intercept is automatically modelled by the nonparametric component. Under this assumption the minimization in (2.3) and (2.4) is unique. For a discussion of these estimates see Severini and Staniswalis (1994).

Our test will be based on a comparison of the semiparametric estimates with the estimators $(\widetilde{\beta}, \widetilde{\gamma})$ in the parametric model

$$(\widetilde{\beta}, \widetilde{\gamma}) = \arg\min_{\beta, \gamma} \mathcal{L}^P(\gamma, \beta). \tag{2.6}$$

Here $\mathcal{L}^P(\gamma, \beta)$ is the quasi-likelihood function in model (1.1)

$$\mathcal{L}^P(\gamma, \beta) = \sum_{i=1}^{n} Q\{G(X_i^T \beta + T_i^T \gamma_i);\, Y_i\}. \tag{2.7}$$

The scale parameter σ can be estimated by

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{\mu}_i)^2/V(\widehat{\mu}_i), \tag{2.8}$$

where $\widehat{\mu}_i = G\{X_i^T\widehat{\beta} + \widehat{m}(T_i)\}$.

A direct comparison of $\widehat{m}(t)$ and $t^T\widetilde{\gamma}$ may be misleading because $\widehat{m}$ has a smoothing bias which is typically nonnegligible. This holds also if the hypothesis of linearity is true. To avoid this effect we will add to $t^T\widetilde{\gamma}$ a bias which will compensate for the bias of $\widehat{m}(t)$. This will be done by the following smoothing step:

$$\widetilde{m} = \arg\min_{m} \int \sum_{i=1}^{n} K_h(t - T_i)\, Q[G\{X_i^T\widetilde{\beta} + m(t)\}; G\{X_i^T\widetilde{\beta} + T_i^T\widetilde{\gamma}\}]\, dt. \tag{2.9}$$

In (2.9) the second argument of $Q(.;.)$ is the parametric estimate of $E(Y_i|X_i, T_i)$ instead of Y_i. The idea behind (2.9) is to apply the smoothing step in (2.3) with $\beta = \widetilde{\beta}$ to the artificial data set $\{G(X_i^T\widetilde{\beta} + T_i^T\widetilde{\gamma}), X_i, T_i\} : i = 1, \ldots, n$.

3 Testing the Parametric versus the Semiparametric Model

Our test procedures are based on the comparison of the parametric estimates $\widetilde{\beta}, \widetilde{m}$ with the semiparametric estimates $\widehat{\beta}, \widehat{m}$. A natural approach would be based on the likelihood ratio statistic $\mathcal{L}(\widehat{m}, \widehat{\beta}) - \mathcal{L}(\widetilde{m}, \widetilde{\beta})$. Unfortunately, this test statistic does not work because in the construction of $\widehat{m}$ and $\widehat{\beta}$ two different likelihood functions (smoothed and unsmoothed) have been used. [A Taylor expansion of the test statistic, in particular of the i-th summand into $c_i\delta_i + d_i\delta_i^2$ with $\delta_i = X_i^T(\widehat{\beta} - \widetilde{\beta}) + \widehat{m}(T_i) - \widetilde{m}(T_i)$, does not lead to a quadratic form.] This cannot be repaired by using the smoothed quasilikelihood $\mathcal{L}^S$ instead of $\mathcal{L}$.

We propose the following test statistic:

$$R_1 = -2\sum_{i=1}^{n} Q(\widetilde{\mu}_i; \widehat{\mu}_i), \tag{3.1}$$

with $\widetilde{\mu}_i = G\{X_i^T\widetilde{\beta} + \widetilde{m}(T_i)\}$ and $\widehat{\mu}_i = G\{X_i^T\widehat{\beta} + \widehat{m}(T_i)\}$ for $i = 1, \ldots, n$.

If the distribution of Y does not belong to an exponential family, the calculation of R_1 involves evaluation of n integrals. In these cases the following two modifications of R_1 are easier to compute. They are motivated by a Taylor expansion of R_1.

$$R_2 = \sum_{i=1}^{n} \frac{[G'\{X_i^T\widehat{\beta} + \widehat{m}(T_i)\}]^2}{V[G\{X_i^T\widehat{\beta} + \widehat{m}(T_i)\}]} \left\{X_i^T(\widehat{\beta} - \widetilde{\beta}) + \widehat{m}(T_i) - \widetilde{m}(T_i)\right\}^2. \tag{3.2}$$

and

$$R_3 = \sum_{i=1}^{n} \frac{\{G'(X_i^T\widetilde{\beta} + T_i^T\widetilde{\gamma})\}^2}{V\{G(X_i^T\widetilde{\beta} + T_i^T\widetilde{\gamma})\}} \left\{X_i^T(\widehat{\beta} - \widetilde{\beta}) + \widehat{m}(T_i) - \widetilde{m}(T_i)\right\}^2 . \qquad (3.3)$$

Theorem 3.1 discusses asymptotics of these test statistics for the case of a binary response Y. The test statistics are asymptotically equivalent on the null hypothesis and have an asymptotic normal distribution.

Theorem 3.1
Suppose that for a model of binary response the assumptions (A1) - (A9) [see Section A2] apply. Then on the hypothesis $m_0(t) = t^T\gamma_0$, it holds that

(i) $R_1 = R_2 + o_p(v_n) = R_3 + o_p(v_n)$,

(ii) $v_n^{-1}(R_1 - e_n) \xrightarrow{D} N(0,1)$,

where $e_n = \lambda \cdot \int K(u)^2\, du (h_1 \cdot \ldots \cdot h_q)^{-1}$, $v_n^2 = 2\lambda \int K^{(2)}(u)^2\, du (h_1 \cdot \ldots \cdot h_q)^{-1}$. Here, λ is the Lebesgue measure of the support S_T of T and $K^{(2)}$ is the convolution of K with itself.

Note in particular, that $\int K(u)^2 \mathrm{d}u \neq \int \{K^{(2)}(u)\}^2 \mathrm{d}u$. Therefore Theorem 3.1 implies that a χ^2 approximation is not appropriate for the distribution of R_1. The reason is that for kernel smoothing operators $\mathcal{K}$ it does not hold that $\mathcal{K}\mathcal{K} = \mathcal{K}$. This is in contrast to projection operators like B-splines, see Buja, Hastie and Tibshirani (1989).

Theorem 3.1 states that the test statistics R_1, R_2 and R_3 are asymptotically equivalent on the hypothesis. By standard arguments of asymptotic decision theory the asymptotic equivalence remains valid for contiguous alternatives (i.e. $n^{-1/2}$ neighbored alternatives). In a parametric setting this would imply that these three tests have asymptotic equivalent power. However, in our nonparametric set up the tests will have nontrivial power (power bounded away from the level and from 1) only for non-contiguous alternatives. Therefore, power functions may behave quite differently. A comparison of power functions based on simulations can be found in the next section.

For two points s_n and t_n the nonparametric estimates $\hat{m}(s_n)$ and $\hat{m}(t_n)$ are asymptotically independent if the support of the kernels $K_h(\bullet - s_n)$ and $K_h(\bullet - t_n)$ are disjoint. This may explain why, asymptotically, R_1 behaves approximately like a sum of $O(h_1^{-1} \cdot \ldots \cdot h_q^{-1})$ independent summands. Because, typically, $h_1^{-1} \cdot \ldots \cdot h_q^{-1}$ is not very large, it can be suspected that normal approximations do not work well for R_1, see Härdle and Mammen (1993) for a related discussion. Therefore, for the calculation of quantiles, we advise not to use normal approximations. Instead, we propose to use the bootstrap and to proceed as follows.

1. Generate samples $(Y_1^*, \ldots, Y_n^*)$ with $E^*(Y_i^*) = G(X_i^T\widetilde{\beta} + T_i^T\widetilde{\gamma})$ and $\mathrm{Var}^*(Y_i^*) = \widehat{\sigma}^2 V\{G(X_i^T\widetilde{\beta} + T_i^T\widetilde{\gamma})\}$. Here E^* and Var^* denote the conditional expectation or variance given $(X_1, T_1, Y_1, \ldots, X_n, T_n, Y_n)$.

2. Calculate estimates $\widehat{\beta}^*, \widehat{m}^*, \widetilde{\beta}^*, \widetilde{\gamma}^*, \widetilde{m}^*$ based on the bootstrap samples (X_1, T_1, Y_1^*), $\ldots$, (X_n, T_n, Y_n^*). Furthermore, calculate test statistics R_1^*, R_2^* and R_3^*. The $(1-\alpha)$ quantiles of the distributions of R_1, R_2, and R_3 can be estimated by the $(1-\alpha)$ quantiles of the conditional distributions of R_1^*, R_2^* or R_3^*, respectively.

In the first step there are a lot of possibilities left for the choice of the conditional distribution of the Y_i^*'s. In the binary response example we considered above, the distribution of Y_i is completely specified by $\mu_i = G(X_i^T\beta + T_i^T\gamma)$. Hence, here it is reasonable to resample from the Bernoulli distribution with parameter $\widetilde{\mu}_i = G(X_i^T\widetilde{\beta} + T_i^T\widetilde{\gamma})$. If the distribution of Y_i cannot be specified (apart from the first two moments) we recommend to use wild bootstrap, see Härdle and Mammen (1993). Theorem 3.2 shows that bootstrap works in case of binary response.

Theorem 3.2
Under the assumptions of Theorem 3.1, it holds for $j = 1, 2, 3$, that

$$d_K(R_j^*, R_j) \xrightarrow{P} 0$$

where d_K denotes the Kolmogorov distance, which is for two probability measures μ and ν (on the real line) defined as

$$d_K(\mu, \nu) = \sup_{t \in \mathbb{R}} \Big| \mu(X \le t) - \nu(X \le t) \Big|.$$

We have seen in our simulations for binary responses that the normal approximation in Theorem 3.1 (ii) is indeed inaccurate for small sample sizes, see Section 4, but that critical values are estimated quite well by bootstrap.

Our test statistic depends on the choice of the bandwidth h. Different values of h may lead to different observed significance levels, see Section 4. Small values of h have been motivated by asymptotic minimax theory , see Ingster (1993) and Lepski and Spokoiny (1995). In particular, the bandwidths proposed in these papers are of smaller order than optimal bandwidths for nonparametric estimation. However, it is difficult to adapt their abstract assumptions to practical settings.

We suggest to apply the test for different choices of h. Differences in observed critical values can be interpreted. Whereas test statistics with small choices of h look more for the appearance of wiggles of small length, large choices of h may detect better global deviances from linearity. So the inspection of the test statistic for different h gives an impression in which respect the function m differs significantly from linear functions.

In case that our test has rejected the hypothesis of linearity it may be of interest to get more insights about the reasons of the rejection. For the case of $d > 1$ we propose to test for average linearity in the direction of one covariate. For a given weight function $w(t_2, \ldots, t_q)$ with $\int w(t_2, \ldots, t_q) dt_2 \cdots dt_q = 1$ we consider the hypothesis that

$$\int m(t_1, \ldots, t_q) w(t_2, \ldots, t_q) dt_2 \cdots dt_q = \alpha t_1 \quad \text{for all } t_1 \text{ and for a scalar } \alpha. \tag{3.4}$$

Testing average linearity of m in t_1 is in particular appropriate in the following model. In this model it is assumed that there is no interaction term of t_1 and $(t_2, \ldots, t_q)$:

$$m(t_1, \ldots, t_q) = m_1(t_1) + m_{2,\ldots,q}(t_2, \ldots, t_q) \quad \text{for some functions } m_1,\ m_{2,\ldots,q}. \tag{3.5}$$

For a discussion of this additive model see Buja et al. (1989) and Hastie and Tibshirani (1990). In this model, hypothesis (3.4) reduces to

$$m_1(t_1) = \alpha t_1 \quad \text{for all } t_1 \text{ and a scalar } \alpha. \tag{3.6}$$

Deviance from average linearity can be measured by the following test statistic

$$R_4 = \min_{a,b} \sum_{i=1}^{n} \frac{[G'\{X_i^T \widehat{\beta} + \widehat{m}(T_i)\}]^2}{V[G\{X_i^T \widehat{\beta} + \widehat{m}(T_i)\}]} \left\{\widehat{m}_1(T_i) - a - bT_i\right\}^2, \tag{3.7}$$

where $\widehat{m}_1(t_1) = \int \widehat{m}(t_1, \ldots, t_q) w(t_2, \ldots, t_q) dt_2 \cdots dt_q$. For the additive model (3.5), the nonparametric estimate $\widehat{m}_1$ of the additive component m_1 has been considered in Linton and Nielsen (1994), Tjøstheim and Auestad (1994), Chen, Härdle, Linton and Severance-Lossin (1996), and Fan, Härdle and Mammen (1995). In a modified definition, the "marginal integration" in the calculation of $\widehat{m}_1$ is replaced by a "marginal summation". For the case of binary response, asymptotics for the estimate $\widehat{m}_1$ is developed in Härdle, Huet, Mammen and Sperlich (1996). Furthermore a proof for asymptotic normality and consistency of bootstrap for the test statistic R_4 can be found there.

4 Simulations and Application

To verify the properties of our test procedure we have run a small simulation study. The following model was used to simulate data from a generalized (partially) linear model

$$E(Y|X = x, T = t) = P(Y = 1|x, t) = \Phi\{\beta x + m(t)\},\ i = 1, \ldots, n,$$

where Φ is the Gaussian distribution function and X, T are independent with uniform distribution on $[-1, 1]$. We performed simulations under the linearity hypothesis using $m(t) = t$. The sample size was $n = 100$ and the number of bootstrap simulations

in each simulated sample $n^* = 250$. Throughout all computations in the paper the Quartic kernel $K(u) = \frac{15}{16}(1-u^2)^2 I(|u| \leq 1)$ was used to generate kernel weights. For the simulations the bandwidth $h = 0.6$ has been used. Table 1 summarizes the results for $m(t) = t$. As can be seen bootstrap seems to work quite accurate for all three test statistics and for all choices of level α.

α	0.01	0.05	0.10	0.15	0.20
R_1	0.012	0.052	0.100	0.152	0.196
R_2	0.012	0.052	0.104	0.144	0.196
R_3	0.008	0.044	0.104	0.140	0.204
	$P(Y = 1\|x,t) = \Phi\{2x + t\}$				

Table 1: Relative number of rejections using the bootstrap method.
$x, t \in [-1, 1]$, 250 Monte Carlo replications, bandwidth $h = 0.6$.

As expected the normal approximation of Theorem 3.1 can be quite inaccurate for this small sample size of $n = 100$ and it should not be used for the calculation of critical values of the test statistics R_1, R_2, R_3. This can be seen from Table 2. At first sight the critical values based on normal approximations are not totally misleading. However, the values in Table 2 concern only the tail of the distributions of R_1, R_2, and R_3 and of the normal limit, given in Theorem 3.1. In the central region there are much larger differences between the distributions of R_1, R_2, and R_3 and the normal limit, given in Theorem 3.1, as can be seen in Figure 2. The normal limit and the distributions of the test statistic are nearly separated. There density estimates for R_1, R_2, R_3 [using the 250 Monte Carlo replications under the linear model $m(t) = t$] are plotted together with the limiting normal density. [The density estimates for R_1, R_2, R_3 are kernel estimates obtained using a bandwidth according to Silverman's rule of thumb: $1.06 \cdot 2.62 \cdot \hat{\sigma} \cdot n^{-1/5}$ for the Quartic kernel. For better comparison, the normal density has been analogously convoluted with a quartic kernel.] Similar plots can be found in Härdle and Mammen (1993) where a related test statistic has been discussed for testing parametric versus nonparametric regression.

α	0.01	0.05	0.10	0.15	0.20
R_1	0.028	0.056	0.076	0.100	0.124
R_2	0.028	0.052	0.064	0.084	0.108
R_3	0.052	0.088	0.112	0.140	0.168
	$P(Y = 1\|x,t) = \Phi\{2x + t\}$				

Table 2: Relative number of rejections using normal approximations.
$x, t \in [-1, 1]$, 250 Monte Carlo replications, bandwidth $h = 0.6$.

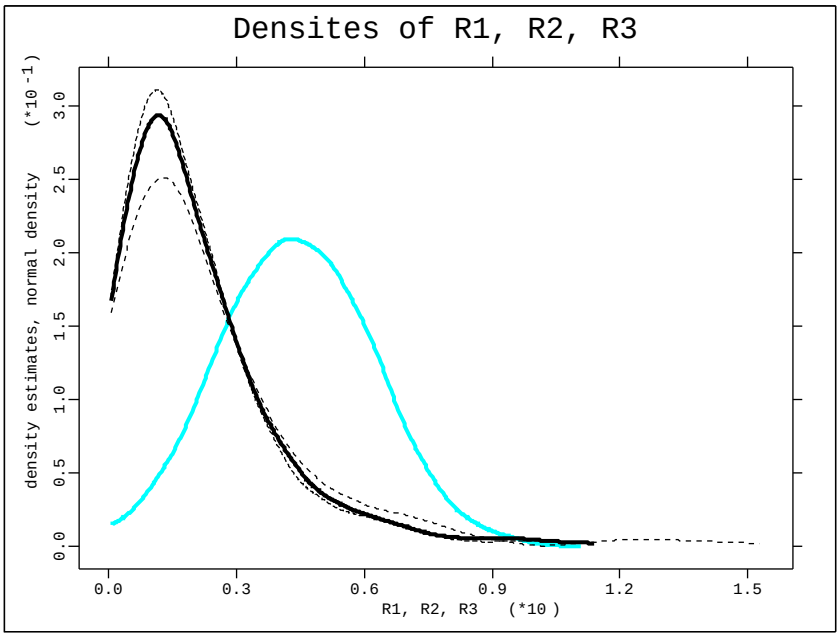

Figure 2: Density estimates for R_1 (thick line), R_2 (thin long dashes), R_3 (thin small dashes) and normal density (thick grey).

Finally we have run our simulations with a function m consisting of a convex combination of the linear model $m(t) = t$ and the nonlinear $m(t) = \cos(\pi t)$. Figure 3 shows the power functions of R_1 for this alternatives (black lines). The power has been plotted for four different significance levels. The power functions for R_2 and R_3 are almost the same and therefore they have been omitted. The grey lines in Figure 3 show (simulated) power functions for a parametric likelihood–ratio test. The hypothesis "$m(x,t) = \Phi\{x\beta + t\gamma\}$ for some β and γ" is tested against the alternative: "$m(x,t) = \Phi\{x\beta + t\gamma + \omega\cos(\pi t)\}$ for some β, γ and ω". Comparison of these two curves gives a first impression on the size of power of our test. The loss of power in the middle region is less than 20% which is not much for an omnibus test.

Let us now return to our introductory example on East–West German migration. Our interest in this subject has been inspired by an analysis of Burda (1993). His paper considers a sample of 3710 East Germans, which have been surveyed in 1991 in the German Socio-Economic Panel, see GSOEP (1991). Among other questions the East German participants have been asked, if they can imagine to move to the Western part of Germany or West Berlin. As in Burda's study we give the value 1 for those who responded positive and 0 if not. The economic model is based on the idea that a person will migrate if its utility (wage differential) will exceed the costs of migration. Of course neither of both variables, wage differential and costs, are directly available. Hence proxy variables need to be used. The original data set of Burda (1993) contains 34 explanatory variables, with four of them continuous (age, income rent, job tenure) and the rest essentially dummy variables (sex, partner, homeowner, family/friends in west, and further variables on occupation, city size, region, education).

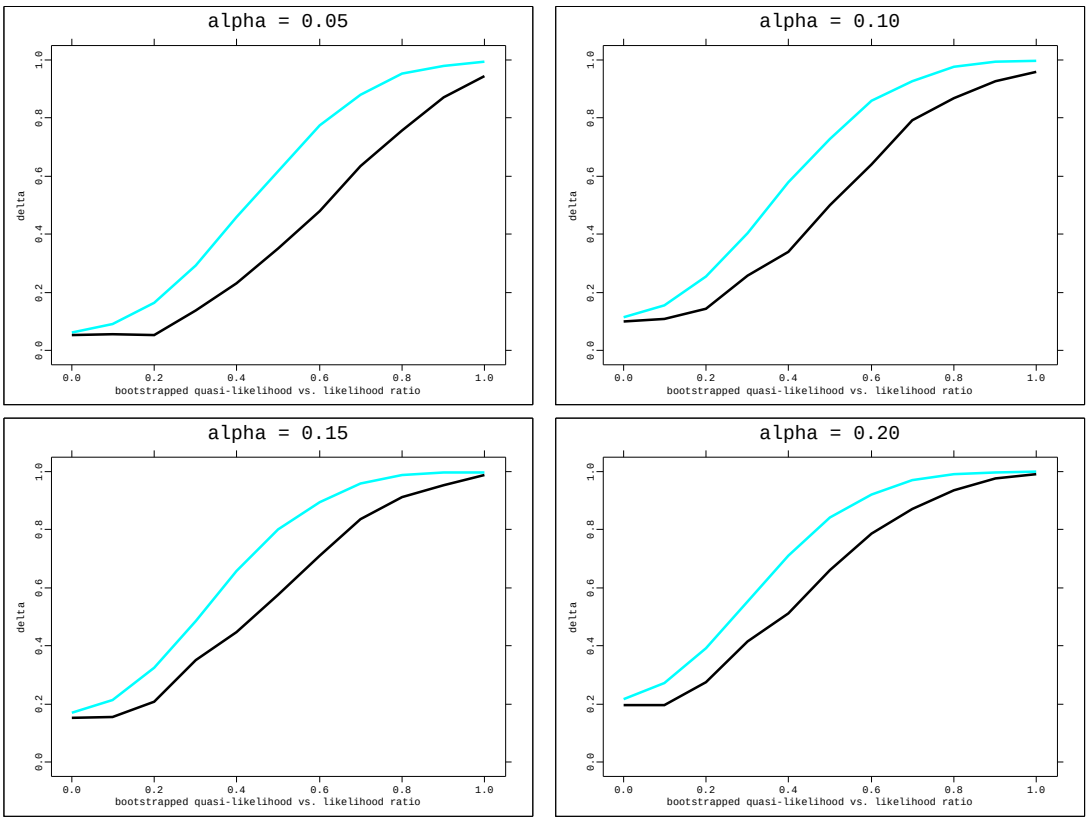

Figure 3: Power functions for $\alpha = 0.05, 0.10, 0.15, 0.20$ (black lines). 250 Monte Carlo Simulations for $P(Y = 1|x, t) = \Phi\{2x+m(t)\}$ with $x, t \in [-1, 1]$ and $m(t) = (1-\nu)t + \nu\cos(\pi t)$, $\nu \in [0, 1]$. Compared to the power of the parametric LR test (grey lines).

		Yes	No	(in %)	
Y	**migration intention**	39.9	60.1		
X_1	**family/friends in west**	88.8	11.2		
X_2	**unemployed/job loss certain**	21.1	78.9		
X_3	**city size 10,000-100,000**	35.8	64.2		
X_4	**female**	50.2	49.8		
		Min	Max	Mean	S.D.
X_5	**age (years)**	18	65	39.93	12.89
T	**houshold income (DM)**	400	4000	2262.22	769.82

Table 3: Descriptive statistics for migration data. Sample from Mecklenburg–Vorpommern, $n = 402$.

It turns out, that regional variables have an important impact on the responses. For instance, the estimation is particularly difficult for East Germans living in East Berlin, since obviously other reasons may influence the intention to migrate than only the wage differential compared to costs. Also, the variables, which are most important, differ

slightly between the five Eastern German states (plus East Berlin). Unemployment, for example, plays a stronger role in the Northern, less industrialized part of East Germany. In the following we give the estimation results for Mecklenburg–Vorpommern (in the very North of Eastern Germany) which leads to a sample size of $n = 402$. We have summarized some descriptive statistics in Table 3.

Table 4 shows the results of a logit fit, using a subset of covariates which have been chosen previously by a model selection procedure based on logit models. For simplicity both continuous variables (age, household income) have been linearly transformed to $[0, 1]$. The migration intention is definitely determined by age. However, also the unemployment, city size and household income variables are highly significant.

	Coeff.	Std.Err.	$P > \|z\|$	Coeff.
const.	-0.358	0.527	0.498	—
family/friends in west	0.589	0.382	0.124	0.599
unemployed/job loss certain	0.780	0.278	0.005	0.800
city size 10,000-100,000	0.822	0.242	0.001	0.842
female	-0.388	0.232	0.094	-0.402
age	-3.364	0.485	< 0.001	-3.329
houshold income	1.084	0.570	0.059	—
	Linear (logit)			Part. Linear

Table 4: Logit coefficients and coefficients in a generalized partially linear model for migration data. Sample from Mecklenburg–Vorpommern, $n = 402$.

A further analysis of this data set by a generalized additive model (keeping the logit link, but generalizing the influence of the age and income variables to nonparametric functions) showed that the age has a nearly perfect linear influence. Because of this relation, we used a generalized partially linear model with a logistic link function and only the influence of household income modelled as a nonparametric function. The coefficients for the parametric covariates are given in Table 4. The resulting fit $\widehat{m}$ (using bandwidth $h = 0.3$) for the function m is that shown in Figure 1 together with the linear fit (thin black dashed line) and the "biased" parametric fit $\widetilde{m}$ (thin grey dashed line). Recall that the estimate $\widetilde{m}$ was an estimate of the sum of the linear function and the bias of $\widehat{m}$, see (2.9).

In Figure 4 we show the functions $\widehat{m}$ and $\widetilde{m}$ (together with the linear fit) for bandwidths $h = 0.1$ and $h = 0.5$. The nonparametric estimate $\widehat{m}$ in the migration example seems to be an obvious nonlinear function. However, it is difficult to judge the significance of the nonlinearity. In general, it cannot be excluded that the difference between the nonparametric and the linear fit may be caused by boundary and bias problems of $\widehat{m}$.

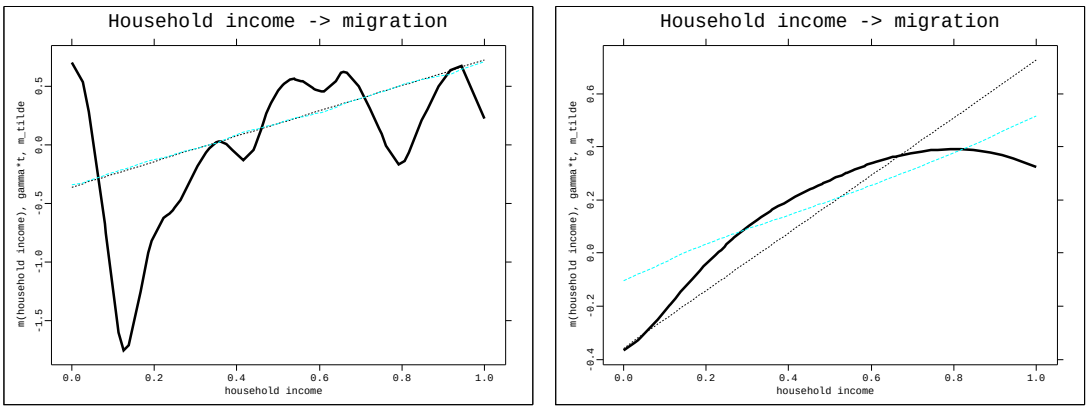

Figure 4: The influence $m(t)$ of household income on migration intention. Nonparametric fit (thick black line), linear fit (thin black dashed line), and "biased" parametric estimate $\widetilde{m}$ (thin grey dashed line). $n = 402$, bandwidths $h = 0.1$ (left) and $h = 0.5$ (right).

h	0.1	0.2	0.3	0.4	0.5
R_1	0.16	0.07	0.05	0.05	0.07
R_2	0.13	0.07	0.05	0.05	0.07
R_3	0.22	0.08	0.05	0.05	0.07

Table 5: Observed significance levels for linearity test for migration data, $n = 402$. 400 bootstrap replications.

Table 5 shows the results of the application of our tests from Section 3. The number of bootstrap simulations is always chosen as $n^* = 400$. We observe that all three tests with R_1, R_2 and R_3 show nearly the same behaviour. The observed significance levels are given for different choices of the bandwidth h. Linearity is rejected (at 5% level) only for bandwidths $0.3, 0.4$. The different behaviour of the test for different h give some indication on possible deviance of m from linear functions. The appearance of wiggles of small length is not significant, see Figure 4 (left panel). However, the global shape of m seems to be not well approximable by linear functions. This result is in accordance with the estimate in Figure 1 and Figure 4 (right panel), where a saturation of the intention to migrate appears for the upper third of the data.

At the end of this section we will shortly present the application of our test statistic in a binary choice regression with a two-dimensional nonparametric function m. The data are a subsample from a training dataset on credit scoring, see Fahrmeir and Tutz (1994) and Fahrmeir and Hamerle (1984). The interest consists in finding how some factors are related to credit worthiness. We used the subsample of loans for cars, which has a sample size of $n = 284$ out of 1000. Some descriptive statistics for this subsample and a selection of covariates can be found in Table 6. The covariate "previous credit o.k." indicates that previous loans were paid without problems or that there were no previous

loans. The variable "employed" takes value 1 if the person taking the loan is employed with the same employer for at least one year. In the following statistical analysis we took logarithms of "amount" and "age" and transformed these values linearly to the interval $[0, 1]$.

		Yes	No	(in %)	
Y	**credit worthy**	73.6	26.4		
X_1	**previous credits o.k.**	66.2	33.8		
X_2	**employed**	73.2	26.8		
		Min	Max	Mean	S.D.
X_4	**duration (months)**	4	54	21.75	10.55
T_1	**amount (DM)**	428	14179	3902.31	2621.95
T_2	**age (years)**	19	75	34.16	10.81

Table 6: Descriptive statistics for credit data. Sample for credits for cars, $n = 284$.

A parametric logit model leads to the parameter estimates listed in Table 7. The influence of employment, duration and amount of credit have the expected sign. The negative influence of "previous credits o.k." is a bit astonishing, but may be explained that also people without previous loan fall in this category. The age variable shows a (global) positive influence in the logit fit, this will change together with the amount variable in the semiparametric fit. Note also, that both coefficients for "amount" and "age" are not significant at 10% level.

	Coeff.	Std.Err.	$P > \|z\|$	Coeff.
const.	2.075	0.616	0.0001	—
previous credits o.k.	-0.698	0.320	0.030	-0.763
employed	0.543	0.311	0.082	0.569
duration	-1.821	0.876	0.039	-2.248
amount	-1.002	1.014	0.324	—
age	0.821	0.688	0.234	—
	Linear (logit)			Part. Linear

Table 7: Logit coefficients and coefficient in partially linear fit for credit scoring, $n = 284$.

In a next step we fitted a generalized partially linear model to the data. Influence of "amount" and "age" has been fitted nonparametrically. The other variables have been modelled as linear covariates. For "duration" this has been done because, typically, it is divisible by 6 months. Figure 5 shows a scatterplot of the two variables "amount" and

"age" on the left panel and the two–variate estimate $\widehat{m}$ (using a bandwidth $h = 0.4$ in both dimensions) on the right panel. It is difficult to check $\widehat{m}$ graphically for significant deviances from linearity. The big peak of $\widehat{m}$ is caused by only a few observations [as can be seen from the scatterplot]. For a closer inspection of $\widehat{m}$ Figure 6 shows the influence of "amount" and "age" separately. In both plots of Figure 6 one variable is held fixed at levels 0.4 (short dashes), 0.5 (thick line) and 0.6 (long dashes). For "age" these levels correspond to 32.9, 37.75, and 43.30 years, respectively. For credit amounts the corresponding original values are DM 1735.90, DM 2463.46, and DM 3495.95, respectively. So obviously, a higher amount of credit seems to get more risky in conjunction with higher age. Also, younger people seem to get less risky with increasing credit amount. Both of these possible conclusions could not be seen from the parametric logit fit.

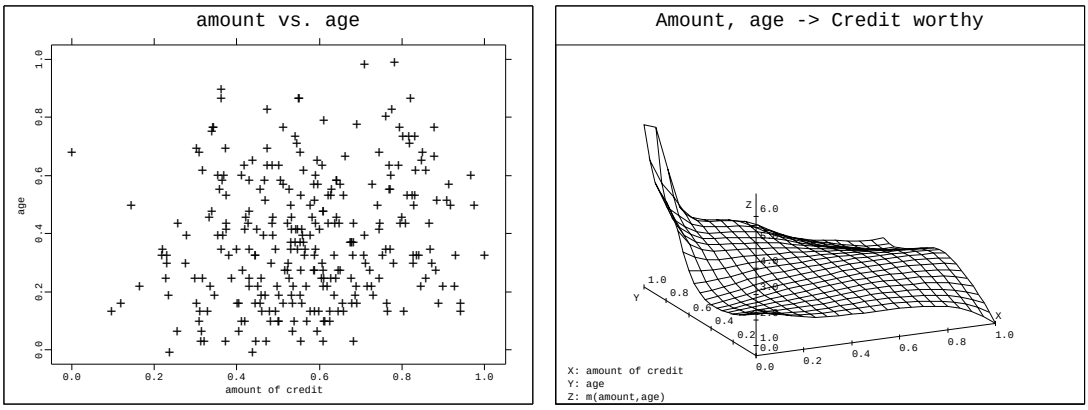

Figure 5: Scatterplot for amount of credit and age (left panel). Influence $\widehat{m}(t_1, t_2)$ of amount and age on credit worthiness (right panel), $n = 284$.

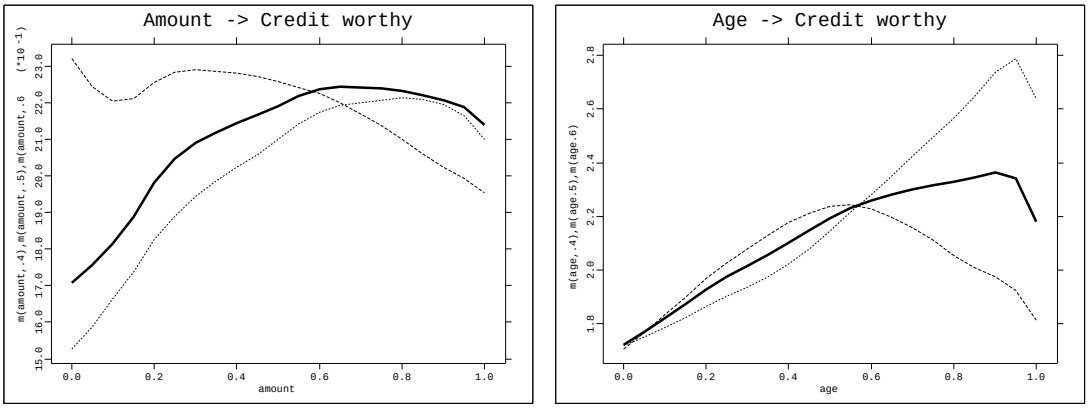

Figure 6: Influence of amount on credit worthiness for fixed age (left panel). Influence of age on credit worthiness for fixed amount (right panel). $n = 284$.

Table 7 gives the observed significance levels of our test statistics for the credit data. Linearity is rejected with high significance only for small values of α. This suggests that deviances from linearity are more locally concentrated.

h	0.2	0.3	0.4	0.5	0.6
R_1	0.04	0.08	0.08	0.09	0.28
R_2	0.02	0.07	0.07	0.09	0.29
R_3	0.24	0.12	0.07	0.08	0.27

Table 8: Observed significance levels for linearity test for credit scoring, $n = 284$. 400 bootstrap replications.

A1 Computational Remarks

In this section we indicate how the estimates in (2.3) and (2.4) can be numerically computed in a binary response model. The following algorithm corresponds to that proposed in Severini and Staniswalis (1994), Example 3, for the special case of a logistic link function.

Put $\eta_j(\beta) = \widehat{m}_\beta(T_j)$ and

$$L_i(u) = Q\{G(u); Y_i\}. \tag{A1.1}$$

Note, that in a binary response model we have

$$\begin{aligned}
L_i(u) &= Y_i \log G(u) + (1 - Y_i) \log\{1 - G(u)\}, \\
L_i'(u) &= \frac{Y_i - G(u)}{G(u)\{1 - G(u)\}} G'(u), \\
L_i''(u) &= \{Y_i - G(u)\} \left[\frac{G''(u)}{G(u)\{1 - G(u)\}} - \frac{\{1 - 2G(u)\}G'(u)^2}{G(u)^2\{1 - G(u)\}^2} \right] - \frac{G'(u)^2}{G(u)\{1 - G(u)\}}.
\end{aligned}$$

Then maximizing the smoothed quasi–likelihood (2.2) requires to solve

$$0 = \sum_{i=1}^{n} L_i'\{X_i^T \beta + \eta_j(\beta)\} K_h(T_i - T_j). \tag{A1.2}$$

Differentiation of (A1.2) leads to $0 = \sum_{i=1}^{n} L_i''\{X_i^T \beta + \eta_j(\beta)\} K_h(T_i - T_j)\{X_i + \eta_j'(\beta)\}$. This gives

$$\eta_j'(\beta) = \frac{-\sum_{i=1}^{n} L_i''\{X_i^T \beta + \eta_j(\beta)\} K_h(T_i - T_j) X_i}{\sum_{i=1}^{n} L_i''\{X_i^T \beta + \eta_j(\beta)\} K_h(T_i - T_j)}. \tag{A1.3}$$

For $\beta = \widehat{\beta}$ it holds that

$$0 = \sum_{i=1}^{n} L_i'\{X_i^T \beta + \eta_i(\beta)\} \{X_i + \eta_i'(\beta)\}. \tag{A1.4}$$

Equations (A1.2), (A1.3), (A1.4) suggest the following iterative Newton–Raphson type algorithm to find $\widehat{\beta}$ and $\widehat{m}(T_j)$, $j = 1, \ldots, n$.

- Start with $\widehat{\beta}^0 = \widetilde{\beta}$, $\widehat{\eta}_j^0 = T_j^T \widetilde{\gamma}$.
- The iteration $k \to k+1$ is determined by the stepwise application of the following two equations:

$$\begin{aligned} 0 &= \sum_{i=1}^{n} L_i'(X_i^T \widehat{\beta}^k + \widehat{\eta}_j^k) K_h(T_i - T_j) + L_i''(X_i^T \widehat{\beta}^k + \widehat{\eta}_j^k) K_h(T_i - T_j)(\widehat{\eta}_j^{k+1} - \widehat{\eta}_j^k) \\ 0 &= \sum_{i=1}^{n} L_i'(X_i^T \widehat{\beta}^k + \widehat{\eta}_i^{k+1}) \widetilde{X}_i^k + L_i''(X_i^T \widehat{\beta}^k + \widehat{\eta}_i^{k+1}) \widetilde{X}_i^k \widetilde{X}_i^{k^T} (\widetilde{\beta}^{k+1} - \widetilde{\beta}^k), \end{aligned}$$

  where

$$\widetilde{X}_j^k = X_j - \frac{\sum_{i=1}^n L_i''(X_i^T \widehat{\beta}^k + \widehat{\eta}_j^{k+1}) K_h(T_i - T_j) X_i}{\sum_{i=1}^n L_i''(X_i^T \widehat{\beta}^k + \widehat{\eta}_j^{k+1}) K_h(T_i - T_j)}.$$

  Then $\widehat{m}^k(T_j) = \widehat{\eta}_j^k$.

Alternatively, the functions $L_i''(u)$ can be replaced by their expectations $-G'(u)^2/V\{G(u)\}$ to obtain a Fisher scoring type procedure.

A2 Assumptions

We state now the assumptions used in the results in Section 3. In the following, the underlying parameters are denoted by β_0, γ_0 and m_0. In the setup of binary responses, the scale parameter σ is equal to 1. We use the notation

$$\begin{aligned} h_{max} &= max\{h_1, \ldots, h_q\}, \\ h_{prod} &= h_1 \cdot \ldots \cdot h_q, \\ \rho &= h_{max}^2 + (n h_{prod})^{-1/2}, \\ \tau &= h_{max} + (n h_{prod})^{-1/2}. \end{aligned}$$

For the asymptotic expansions we make the following assumptions.

(A1) $(X_1, T_1, Y_1), \ldots, (X_n, T_n, Y_n)$ are i.i.d. tuples. T_i takes values in $\mathbb{R}^q$, X_i is $\mathbb{R}^p$ valued, and Y_i is $\{0, 1\}$ valued.

(A2) $E(Y_i | X_i, T_i) = G\{X_i^T \beta_0 + m_0(T_i)\}$ with $\beta_0 \in \mathbb{R}^p$.

(A3) $X_i^T \beta_0 + m_0(T_i)$ has compact support S. X_i and T_i have compact convex support S_X, S_T. T_i has a twice continuously differentiable density f_T with $\inf_{t \in S_T} f_T(t) > 0$.

(A4) There exists an $\varepsilon > 0$ such that

$$G(u)^{-1}, \{1 - G(u)\}^{-1}, G^{(k)}(u), \quad k = 1, \ldots 4,$$

are bounded on $u \in S^\varepsilon = \{v : \exists v' \in S \text{ with } |v' - v| \leq \varepsilon\}$.

(A5) m is twice continuously differentiable on $\mathbb{R}$.

(A6) The kernel K is a product kernel $K(u) = K_1(u_1) \cdot \ldots \cdot K_q(u_q)$. The kernels K_j are symmetric probability densities with compact support ($[-1,1]$, say), $j = 1, \ldots, q$.

(A7) The estimate $\widehat{\beta}$ is defined as $\arg \min_{\beta: \|\beta - \beta_0\| \leq \rho} \mathcal{L}(\widehat{m}_\beta, \beta)$. For a δ_n with $\delta_n \to 0$ the estimate $\widehat{m}_\beta(t)$ is defined as $\arg \min_{\eta: |\eta - m_0(t)| \leq \delta_n} \sum_{i=1}^{n} L_i(X_i^T \beta + \eta) K_h(T_i - t)$.

(A8) $E\left[L_1''\{X_1^T \beta_0 + m_0(T_1)\} | T_1 = t\right]$ and $E\left[L_1''\{X_1^T \beta_0 + m_0(T_1)\} X_1 | T_1 = t\right]$ are twice continuously differentiable functions for $t \in S_T$.

(A9) $h_{prod}\, n^{1/2} (\log n)^{-1} \to \infty$ and $h_{max} = o(n^{-1/8} (\log n)^{-1/4})$.

(A10) m is four times continuously differentiable on $\mathbb{R}$. The support S_T is of the form $S_T^1 \times S_T^2$ with $S_T^1 \subset \mathbb{R}$ and $S_T^2 \subset \mathbb{R}^{q-1}$. The weight function w is positive and twice continuously differentiable. Furthermore, for a $\delta > 0$ we have that $w(t) = 0$ for $t \in \{s : \text{ there exists an } u \notin S_T^2 \text{ with } \|s - u\| \leq \delta\}$.

(A11) $h_1 = o(n^{-1/4} (\log n)^{-1/4})$.

A3 Proofs

In this section we always assume that (A1) - (A8) hold. The following lemmas give the stochastic expansions for $\widehat{\beta}$ and $\widehat{m}$. Recall that the set S_T was the (compact) support of T_i. We denote $S_T^- = \{t \in S_T : t + \eta \in S_T \text{ for all } \eta \text{ with } |\eta_j| \leq h_j (j = 1, \ldots, q\}$ and $S_T^h = S_T \setminus S_T^-$. Furthermore, define

$$
\begin{aligned}
S_{i,1} &= L_i'\{X_i^T \beta_0 + m_0(T_i)\}, \quad S_{i,2} = L_i''\{X_i^T \beta_0 + m_0(T_i)\}, \\
\widetilde{X}_i &= X_i - \{E[S_{i,2} | T_i]\}^{-1} E[S_{i,2} X_i | T_i], \\
w_i(t) &= K_h(t - T_i) \Big\{ n^{-1} \sum_{j=1}^{n} K_h(t - T_j) \Big\}^{-1}.
\end{aligned}
$$

Lemma A3.1

(i) For all $C > 0$ it holds that

$$
\sup_{\substack{t \in S_T^- \\ \|\beta - \beta_0\| \leq C\rho}} \Bigg| \widehat{m}_\beta(t) - \Bigg(m(t) - \{E(S_{1,2} | T_1 = t)\}^{-1} \Big[\frac{1}{n} \sum_{i=1}^{n} w_i(t) L_i'\{X_i^T \beta_0 + m_0(t)\} + E(S_{1,2} X_1^T | T_1 = t)(\beta - \beta_0) \Big] \Bigg) \Bigg| = O_p(\rho^2 \log n).
$$

(ii) The supremum in (i) taken over $t \in S_T^h, \|\beta - \beta_0\| \leq C\rho$ *is of stochastic order* $O_p(\tau^2)$.

Proof

We prove only statement (i). Choose $C > 0$. We have for $t \in S_T^-, \|\beta - \beta_0\| \leq C\rho$

$$\sum_{i=1}^{n} L_i'\{X_i^T\beta + \widehat{m}_\beta(t)\}K_h(t - T_i) = 0. \tag{A3.1}$$

This follows from

$$\sup \sum_{i=1}^{n} L_i''(X_i^T\beta + \eta)K_h(t - T_i) < 0 \tag{A3.2}$$

with probability tending to one, where the supremum runs over $|\eta - m_0(t)| \leq \delta_n, t \in S_T^-$, and β with $\|\beta - \beta_0\| \leq C\rho$.

Note that (A3.2) implies that, if we find an $\eta_\beta(t)$ with $|\eta_\beta(t) - m_0(t)| \leq \delta_n$ and

$$\sum_{i=1}^{n} L_i'\{X_i^T\beta + \eta_\beta(t)\}K_h(t - T_i) = 0,$$

then with probability tending (uniformly) to one we get $\widehat{m}_\beta(t) = \eta_\beta(t)$. Inequality (A3.2) can be shown by using the boundedness of L_i''' and

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} L_i''\{X_i^T\beta_0 + m_0(T_i)\} \\ &\qquad \xrightarrow{P} -E\left(\frac{G'\{X_1^T\beta_0 + m_0(T_1)\}^2}{G\{X_1^T\beta_0 + m_0(T_1)\}[1 - G\{X_1^T\beta_0 + m_0(T_1)\}]}\right), \end{aligned} \tag{A3.3}$$

see Assumption (A7). Equation (A3.1) implies

$$\begin{aligned} 0 \;=\; &\frac{1}{n}\sum_{i=1}^{n} w_i(t)L_i'\{X_i^T\beta_0 + m_0(t)\} \\ &+\frac{1}{n}\sum_{i=1}^{n} w_i(t)L_i''\{X_i^T\beta_0 + m_0(t)\}\ \{\widehat{m}_\beta(t) - m_0(t) + X_i^T(\beta - \beta_0)\} \\ &+R_1(\beta, t)\left[\{\widehat{m}_\beta(t) - m_0(t)\}^2 + \rho^2\right] \end{aligned} \tag{A3.4}$$

with

$$\sup_{\substack{t \in S_T^- \\ \|\beta - \beta_0\| \leq C\rho}} |R_1(\beta, t)| \leq C_1 \quad \text{(a.s)}$$

for a constant $C_1 > 0$ for n large enough. Furthermore, we have $|\widehat{m}_\beta(t) - m_0(t)| \leq \delta_n \to 0$, see (A7). This implies

$$\begin{aligned} \widehat{m}_\beta(t) \;=\; &m_0(t) - \left[\frac{1}{n}\sum_{i=1}^{n} w_i(t)L_i''\{X_i^T\beta_0 + m_0(t)\}\right]^{-1} \\ &\left[\frac{1}{n}\sum_{i=1}^{n} w_i(t)L_i'\{X_i^T\beta_0 + m_0(t)\} + \frac{1}{n}\sum_{i=1}^{n} w_i(t)L_i''\{X_i^T\beta_0 + m_0(t)\}X_i^T(\beta - \beta_0)\right] \\ &+R_2(\beta, t)\,\rho^2 \log n, \end{aligned} \tag{A3.5}$$

where

$$\sup_{\substack{t\in S_T^- \\ \|\beta-\beta_0\|\le C\rho}} |R_2(\beta,t)| = O_p(1).$$

For (A3.5) it has been used that

$$\sup_{t\in S_T^-}\left|\frac{1}{n}\sum_{i=1}^n w_i(t)L_i'\{X_i^T\beta_0+m_0(t)\}\right| = O_p(\rho\sqrt{\log n}).$$

This follows from

$$\sup_{t\in S_T^-}\left|\frac{1}{n}\sum_{i=1}^n K_h(t-T_i)\left[L_i'\{X_i^T\beta_0+m_0(t)\}-L_i'\{X_i^T\beta_0+m_0(T_i)\}\right]\right| = O_p(\rho)$$

and

$$\sup_{t\in S_T^-}\left|\frac{1}{n}\sum_{i=1}^n K_h(t-T_i)L_i'\{X_i^T\beta_0+m_0(T_i)\}\right| = O_p(\rho\sqrt{\log n}).$$

Recall that $E\left[L_i'\{X_i^T\beta_0+m_0(T_i)\}|X_i,T_i\right]=0$. For the statement of the lemma it remains to show

$$\sup_{t\in S_T^-}\left|\frac{1}{n}\sum_{i=1}^n w_i(t)L_i''\{X_i^T\beta_0+m_0(t)\}-E(S_{1,2}|T_1=t)\right| = O_p(\rho\sqrt{\log n}) \tag{A3.6}$$

$$\sup_{t\in S_T^-}\left\|\frac{1}{n}\sum_{i=1}^n w_i(t)L_i''\{X_i^T\beta_0+m_0(t)\}X_i^T-E(S_{1,2}X_1^T|T_1=t)\right\| = O_p(\rho\sqrt{\log n}). \tag{A3.7}$$

For the proof of (A3.6) note first that

$$\sup_{t\in S_T^-}\left|\frac{1}{n}\sum_{i=1}^n w_i(t)\left[L_i''\{X_i^T\beta_0+m_0(t)\}-L_i''\{X_i^T\beta_0+m_0(T_i)\}\right]\right| = O_p(\rho),$$

see (A4). With the help of (A8) one shows

$$\sup_{t\in S_T^-}\left|\frac{1}{n}\sum_{i=1}^n w_i(t)L_i''\{X_i^T\beta_0+m_0(T_i)\}-E(S_{1,2}|T_1=t)\right| = O_p(\rho\sqrt{\log n}).$$

Equation (A3.7) can be shown similarly.

Lemma A3.2

(i) For all $C>0$ it holds that

$$\sup_{\substack{t\in S_T^- \\ \|\beta-\beta_0\|\le C\rho}}\left\|\frac{\partial\widehat{m}_\beta(t)}{\partial\beta}+\{E(S_{1,2}|T_1=t)\}^{-1}E(S_{1,2}X_1|T_1=t)\right\| = O_p(\rho\sqrt{\log n}).$$

(ii) The supremum in (i) taken over $t \in S_T^h, \|\beta - \beta_0\| \leq C\rho$ *is of stochastic order* $O_p(\tau)$.

Proof

Lemma A3.2 can be proved similarly as Lemma A3.1. One uses that

$$\sum_{i=1}^{n} L_i''\{X_i^T\beta + \widehat{m}_\beta(t)\}K_h(t-T_i)\frac{\partial}{\partial\beta}\widehat{m}_\beta(t) \tag{A3.8}$$
$$+\sum_{i=1}^{n} L_i''\{X_i^T\beta + \widehat{m}_\beta(t)\}X_iK_h(t-T_i) = 0.$$

Lemma A3.3

For the estimate $\widehat{\beta}$ *the following stochastic expansion holds*

$$\widehat{\beta} = \beta_0 + \{E(S_{1,2}\widetilde{X}_1\widetilde{X}_1^T)\}^{-1}\frac{1}{n}\sum_{i=1}^{n} S_{i,1}\widetilde{X}_i + O_p(\rho^2\log n).$$

Proof

We show that with probability tending to one there exists a solution β with $\|\beta-\beta_0\| \leq \rho$ of the following equation and that (with probability tending to one) this solution is unique.

$$\frac{\partial}{\partial\beta}\sum_{i=1}^{n} L_i\{X_i^T\beta + \widehat{m}_\beta(T_i)\} = 0. \tag{A3.9}$$

Expansion of the left hand side of (A3.9) gives with the help of Lemma A3.2

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} L_i'\{X_i^T\beta + \widehat{m}_\beta(T_i)\}\left[X_i + \frac{\partial}{\partial\beta}\widehat{m}_\beta(T_i)\right] \qquad \text{(A3.10)}\\ &= \frac{1}{n}\sum_{i=1}^{n} L_i'\{X_i^T\beta_0 + m_0(T_i)\}\left[X_i + \frac{\partial}{\partial\beta}\widehat{m}_\beta(T_i)\right]\\ &\quad +\frac{1}{n}\sum_{i=1}^{n} L_i''\{X_i^T\beta_0 + m_0(T_i)\}\widetilde{X}_iX_i^T(\beta-\beta_0)\\ &\quad +\frac{1}{n}\sum_{i=1}^{n} L_i''\{X_i^T\beta_0 + m_0(T_i)\}\widetilde{X}_i\left[\widehat{m}_\beta(T_i) - m_0(T_i)\right] + O_p(\rho^2\log n). \end{aligned}$$

This expansion holds uniformly for β with $\|\beta-\beta_o\| \leq \rho$. For instance, it has been used that

$$\sup_{\substack{t\in S_t^- \\ \|\beta-\beta_0\|\leq\rho}} |\widehat{m}_\beta(t) - m(t)| = O_p(\rho\sqrt{\log n}).$$

This follows by standard techniques from Lemma A3.1. By expansion of (A3.8) it can be shown that

$$\frac{1}{n}\sum_{i=1}^{n} L_i'\{X_i\beta_0 + m_0(T_i)\}\left[X_i + \frac{\partial}{\partial\beta}\widehat{m}_\beta(T_i)\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n} L_i'\{X_i^T\beta_0 + m_0(T_i)\}\widetilde{X}_i + O_p(\rho^2).$$

Plugging this into the right hand side of (A3.10) and replacing averages by their expectations gives that (with probability tending to one) there exists a solution $\beta = \overline{\beta}$ of (A3.9) with

$$\overline{\beta} = \beta_0 + \{E(S_{1,2}\widetilde{X}_1\widetilde{X}_1^T)\}^{-1}\,\frac{1}{n}\sum_{i=1}^{n} S_{i,1}\widetilde{X}_i + \; O_p(\rho^2 \log n).$$

Because of $\overline{\beta} - \beta_0 = O_p(n^{-1/2})$, we have $\overline{\beta} = \widehat{\beta}$ (with probability tending to one). This shows Lemma A3.3.

With the help of Lemmas A3.1 and A3.2 we get for the estimate $\widehat{m}$ the following expansion.

Corollary A3.4

(i) For the estimate $\widehat{m}$ the following stochastic expansion holds:

$$\sup_{t\in S_T^-}\Big|\widehat{m}(t) - \{\overline{m}(t) + \{E(S_{1,2}|T_1 = t)\}^{-1}E(S_{1,2}X_1^T|T_1 = t)$$
$$\{(S_{1,2}\widetilde{X}_1\widetilde{X}_1)\}^{-1}\frac{1}{n}\sum_{i=1}^{n} S_{i,1}\widetilde{X}_i\}\Big| = \; O_p(\rho^2\sqrt{\log n}),$$

with $\overline{m}(t) = m_0(t) + E(S_{1,2}|T_1 = t)^{-1}\frac{1}{n}\sum_{i=1}^{n} w_i(t)L_i'\{X_i^T\beta_0 + m_0(t)\}$.

(ii) The supremum in (i) taken over $t \in S_T^h$ is of stochastic order $O_p(\tau^2)$.

In particular, we get $\sup_{t\in S_T^-}|\widehat{m}(t) - \overline{m}(t)| = O_p(n^{-1/2})$ and $\sup_{t\in S_T^h}|\widehat{m}(t) - \overline{m}(t)| = O_p(\tau^2)$. Also $\sup_{t\in S_T^-}|\widehat{m}(t) - m(t)| = O_p(\rho\sqrt{\log n})$ and $\sup_{t\in S_T^h}|\widehat{m}(t) - m(t)| = O_p(\tau)$.

In Section 2 we introduced in (2.9) the modification $\widetilde{m}(t)$ of the parametric estimate $t^T\widetilde{\gamma}$. The purpose of this modification was to compensate for the bias of $\widehat{m}(t)$ when comparing $\widetilde{m}(t)$ and $\widehat{m}(t)$. The next lemma shows that this modification works.

Lemma A3.5
Suppose that the hypothesis (1.1) holds, i.e. $m_0(t) = t^T\gamma_0$.

$$\sup_{t\in S_T^-}\Big|\widetilde{m}(t) - t^T(\widetilde{\gamma} - \gamma_0) - E\{\overline{m}(t)|X_1, T_1, \ldots, X_n, T_n\}\Big| = \; O_p(\rho^2\sqrt{\log n}).$$

Proof
The proof uses similar expansions as above. In particular it uses the fact that with probability tending to one

$$\sum_{i=1}^{n} K_h(t - T_i)\frac{G\{\widetilde{\mu}_i(t)\} - G(X_i^T\widetilde{\beta} + T_i^T\widetilde{\gamma})}{G\{\widetilde{\mu}_i(t)\}[1 - G\{\widetilde{\mu}_i(t)\}]}\,G'\{\widetilde{\mu}_i(t)\} = 0,$$

where $\widetilde{\mu}_i(t) = X_i^T \widetilde{\beta} + \widetilde{m}(t)$.

Proof of Theorem 3.1

Application of the foregoing expansions for the parametric and semiparametric estimates gives:

$$\sup_{t \in S_T^-} \left| [\widehat{m}(t) - \widetilde{m}(t)] - [\overline{m}(t) - E\{\overline{m}(t) | X_1, T_1, \ldots, X_n, T_n\}] \right| = O_p(\rho^2 \sqrt{\log n}),$$

$$\sup_{t \in S_T^-} \left| \overline{m}(t) - E\{\overline{m}(t) | X_1, T_1, \ldots, X_n, T_n\} \right| = O_p((nh_{prod})^{-1/2} \sqrt{\log n}),$$

These equalities together with the expansions for the suprema over S_T^- imply for $j = 1, 2, 3$

$$\begin{aligned} R_j &= R + O_p(n\rho^2 (nh_{prod})^{-1/2} \log n), \\ R &= \sum_{i=1}^{n} \frac{G'(\eta_i)^2}{G(\eta_i)\{1 - G(\eta_i)\}} \{\overline{m}(T_i) - E[\overline{m}(T_i) | X_1, T_1, \ldots, X_n, T_n]\}^2, \end{aligned}$$

where $\eta_i = X_i^T \beta_0 + T_i^T \gamma_0$ for $i = 1, \ldots, n$. Under our assumptions, we have $n\rho^2 (nh_{prod})^{-1/2} \log n = o(h_{prod}^{-1/2}) = o(v_n)$. This shows statement (i). For statement (ii) note that, conditionally given $X_1, T_1, \ldots, X_n, T_n$, the statistic R is a U-statistic. Proceeding as in Härdle and Mammen (1993) one can verify de Jong's (1987) conditions for asymptotic normality of U-statistics.

Proof of Theorem 3.2

As in the proof of Theorem 3.1 one shows for $j = 1, 2, 3$ that

$$d_K\{R_j^*, N(e_n, v_n^2)\} \longrightarrow 0 \quad \text{(in probability)}.$$

(Recall that e_n and v_n have been introduced in Theorem 3.1.)

References

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion), *Annals of Statistics* **17**: 453–555.

Burda, M. (1993). The determinants of east–west german migration, *European Economic Review* **37**: 452–461.

Carroll, R., Fan, J., Gijbels, I. and Wand, M. (1995). Generalized partially single-index models, *Technical report*, Department of Statistics, Texas A&M University.

Chen, R., Härdle, W., Linton, O. and Severance-Lossin, E. (1996). Estimation and variable selection in additive nonparametric regression models, *in* W. Härdle and M. Schimek (eds), *Proceedings of the COMPSTAT Satellite Meeting Semmering 1994*, Physica Verlag, Heidelberg.

de Jong, P. (1987). A central limit theorem for generalized quadratic forms, *Probability Theory and Related Fields* **75**: 261–277.

Fahrmeir, L. and Hamerle, A. (1984). *Multivariate Statistische Verfahren*, De Gruyter, Berlin.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer.

Fan, J., Härdle, W. and Mammen, E. (1995). Direct estimation of low dimensional components in additive models, *Discussion paper*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models, *International Statistical Review* **55**: 245–259.

GSOEP (1991). *Das Sozio-ökonomische Panel (SOEP) im Jahre 1990/91*, Projektgruppe "Das Sozio-ökonomische Panel", Deutsches Institut für Wirtschaftsforschung. Vierteljahreshefte zur Wirtschaftsforschung, pp. 146–155.

Härdle, W., Huet, S., Mammen, E. and Sperlich, S. (1996). Semiparametric additive indices for binary response, *Technical report*, Sonderforschungsbereich 373, Humboldt-Universität zu Berlin.

Härdle, W. and Mammen, E. (1993). Testing parametric versus nonparametric regression, *Annals of Statistics* **21**: 1926–1947.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I - III, *Math. Methods of Statist.* **2**: 85 – 114, 171 – 189, 249 – 268.

Lepski, O. V. and Spokoiny, V. G. (1995). Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative, unpublished manuscript.

Linton, O. and Nielsen, J. P. (1994). A kernel method of estimating structured nonparametric regression based on marginal integration, *Biometrika*. in press.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*, Econometric Society Monographs No. 4, Cambridge University Press.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Vol. 37 of *Monographs on Statistics and Applied Probability*, 2 edn, Chapman and Hall, London.

Robinson, P. M. (1988). Root n-consistent semiparametric regression, *Econometrica* **56**: 931–954.

Severini, T. A. and Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models, *Journal of the American Statistical Association* **89**: 501–511.

Severini, T. A. and Wong, W. H. (1992). Generalized profile likelihood and conditionally parametric models, *Annals of Statistics* **20**: 1768–1802.

Speckman, P. E. (1983). Regression analysis for partially linear models, *Journal of the Royal Statistical Society, Series B* **50**: 413–436.

Tjøstheim, D. and Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: projections., *Journal of the American Statistical Association* **89**: 1398–1409.